Régression linéaire

Le but d'une régression linéaire est de trouver la meilleure relation affine entre deux séries statistiques.

Motivation

On possède deux séries statistiques de données entre lesquelles on veut trouver ou vérifier une relation. Il est très facile de vérifier qu'une relation est linéaire entre deux séries. En revanche, il est difficile de savoir si une relation est logarithmique, parabolique, ... L'idée est donc de construire des séries de données à partir des deux séries d'origine, et de vérifier que le lien entre ces deux nouvelles séries est linéaire. Il faut toutefois bien sûr avoir une idée de la relation à construire !

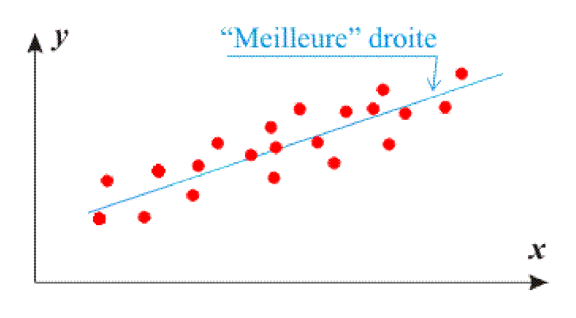
Il y a donc plusieurs étapes :

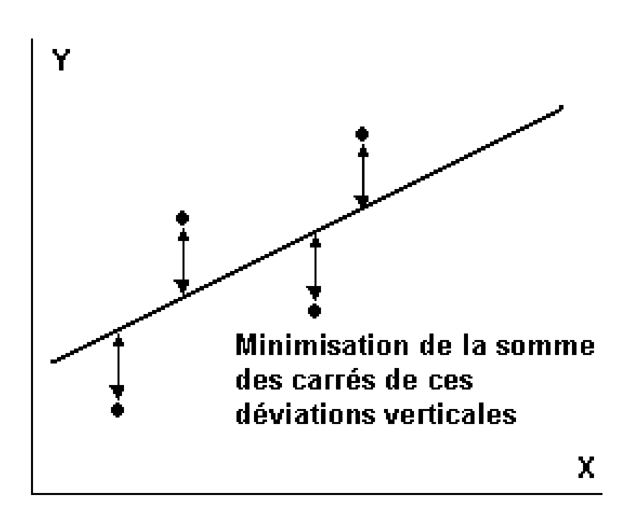
- « deviner » la relation entre les deux séries statistiques, et éventuellement calculer les nouvelles séries nécessaires ;
- effectuer une régression linéaire entre ces deux séries ;
- vérifier si elle est de bonne qualité : si c'est le cas, on pourra considérer que la relation est effectivement correcte ; sinon, c'est que la relation de départ n'était pas bonne !

On va expliquer tout d'abord le principe d'une régression, et dire comment on peut quantifier sa « qualité » ; puis on donnera quelques infos sur les calculatrices. Enfin, on donnera un exemple et une petite liste d'erreurs classiques.

Principe

Si on place dans un plan les points correspondants aux couples formés par les deux séries statistiques fournies, on obtient ce qu'on appelle un nuage de points. Effectuer une régression linéaire entre les deux séries consiste à trouver la droite qui passe au plus près de l'ensemble de ses points. La calculette cherchera ainsi toujours l'équation d'une droite, même si visiblement, cela n'a aucun sens (par exemple, si le lien semble être exponentiel, on obtiendra quand même une équation : elle sera de mauvaise qualité toutefois).





Mais la notion de « meilleure droite » est subjective ! Comment déterminer ce qu'est une bonne droite ? Il y a de nombreux critères existants, le plus courant étant celui des « **moindres carrés** », très souvent utilisé, notamment en SI. Pour cela, on cherche une droite d'équation y=ax+b, où x et y sont les deux séries statistiques, et on évalue l'erreur commise entre le point réel et le point de même abscisse (ou ordonnée) sur la droite. Puisque l'erreur commise est tantôt positive, tantôt négative, et est de plus en général aléatoire, la moyenne de ses erreurs sera souvent nulle : la somme des erreurs est donc une mauvaise idée. L'astuce consiste à retenir comme grandeur **la somme des carrés des erreurs** : on ajoute ainsi des grandeurs positives, et c'est cette grandeur que l'on va chercher à minimiser, d'où l'appellation de **critère des moindres carrés.**

Le calcul mathématique permet alors de calculer trois éléments : la pente de la droite a, son ordonnée à l'origine b, et le coefficient de régression (ou de corrélation) r (compris entre -1 et 1) ou parfois r^2 (compris entre 0 et 1, fatalement). C'est la calculatrice qui effectue cette opération, même si les calculs ne sont en fait pas très compliqués et font intervenir des grandeurs statistiques caractéristiques des deux séries. Mais cela dépasse le cadre du cours.

Remarque : il est possible de chercher des relations plus compliquées entre par exemple plus de deux séries statistiques : on parle alors de régression multiple, mais c'est très au-delà de nos besoins en prépa.

Interprétation des coefficients r et r² : qualité d'une régression

Ça y est, la calculatrice nous a donné une relation affine entre nos deux séries, mais on ne sait pas encore si c'est une droite de bonne qualité (car, rappel, la calculatrice vous trouvera toujours une relation affine, mais celle-ci sera peut-être catastrophique et ne collera pas du tout avec les données) ! La calculette répond à votre question « quelle est la meilleure droite passant dans ce nuage de points », mais elle vous dit également « voilà la droite, mais bon, c'est un peu idiot car ce n'est pas une bonne droite, pour un tel nuage c'est pas terrible » à l'aide du coefficient r (ou r^2). Ce dernier est d'autant plus proche de 1 que la qualité de la régression est élevée.

On considèrera que la droite est de bonne qualité si on a au moins deux ou trois 9 après le 0 pour r (ou r^2). Ainsi, la qualité sera suffisante si $r>0,99$ ou mieux 0,999 (ou $r<-0,99$ voire -0,999) si r est négatif. L'emploi de r^2 lève toute ambiguïté sur ce dernier point.

Une valeur de $r=0,97$ n'est en effet pas suffisante pour dire que la régression est correcte (on fera le test plus loin).

Différents types de calculettes

Chacun doit savoir effectuer une régression avec sa calculatrice, et chacun est responsable de la maîtrise de sa propre calculatrice. Voici quelques infos, mais bien sûr non exhaustives (y compris pour chaque marque)... En cas de besoin, lisez votre mode d'emploi.

- **Casio :** On va dans le menu STAT, et on rentre les deux séries statistiques dans une liste pour chaque série. On peut déjà regarder le nuage en tapant GRAPH puis en choisissant un (GPH1 par exemple). Apparaît alors sous le graph un menu : si on veut une régression en X, on appuie sur la case X justement. Une fenêtre LinearReg donne alors les valeurs de a, b et r, en précisant l'équation tracée. Attention ! Une erreur classique est de ne pas avoir dit entre quelles séries on effectuait la régression. Cela se règle lorsque l'on appuie sur GRAPH, et qu'on choisit le bouton SET, qui permet de dire quelle liste joue le rôle de « x » et laquelle joue le rôle des « y ».
- **TI :** Ouvrir l'éditeur de données via [STAT] [EDIT] [1]. On arrive dans l'éditeur, le curseur étant positionné sur la cellule L1(1). Il suffit de taper les valeurs dans les différentes cellules : les valeurs de x dans la colonne L1 et celles de y dans L2. Ouvrir l'éditeur de graphes avec [2nd] [STAT PLOT] [1]. Valider l'option [On] puis choisir le type de points, par exemple le nuage (tracé discontinu). Entrer les listes correspondant aux abscisses et aux ordonnées : L1et L2. Choisir le type de marque, par exemple la boîte (carré). Afficher automatiquement les points sur la totalité de l'écran en réglant le zoom : [ZOOM] [9]. Revenir dans l'éditeur de données avec [STAT] et choisir le mode calcul par [CALC]. Si les points sont à peu près alignés, on choisit comme modèle la régression linéaire [5:LinReg(ax+b)], et on précise abscisses et ordonnées : L1, L2. Pour visualiser la droite de régression, il faut passer par [Y][CLEAR] puis [VARS] [5:Statistics...] et sélectionner [EQ] [7:RegEQ]. Le retour à l'écran graphique par [GRAPH] permet de retrouver les points expérimentaux et la droite de régression : il est possible de se déplacer sur chacune des représentations grâce à [TRACE] associée aux touches du curseur.

Exemple 1 : quand tout va bien

Temps (s)	10	15	23	38	45	49	56	62	70
Hauteur (m)	20,618	31,307	48,393	80,428	95,370	103,900	118,823	131,660	148,723

Après avoir rentré les données, on peut déjà voir que le nuage ressemble à une droite. La régression donne Hauteur = 2,1.Temps-0,7 avec un coefficient de régression r=0,999... La droite est donc de bonne qualité et la régression validée.

Remarque : Inutile de donner trop de chiffres significatifs, puisque de toutes façons la droite est une approximation du nuage de points ... Deux (ou trois) CS seront donc largement suffisants.

Exemple 2 : tout n'est pas linéaire !

t (s)	0	1,2	3	6	9	12	15	18	24
$[I_2]$ (mmol.ℓ^{-1})	20	12	10	5,4	2,8	1,4	0,75	0,4	0,1

Ici, le nuage n'est visiblement pas linéaire ! Il ressemble plus à un nuage exponentiel. On va donc faire l'hypothèse que la concentration en diiode est une fonction exponentielle du temps. On aurait donc une relation du type

$$[I_2] = a.e^{-bt}$$

Comme on cherche une droite, on peut évaluer le logarithme de la concentration en diiode puisque

$$\ln [I_2] = \ln a - b.t$$

qui est bien une relation linéaire (si toutefois notre hypothèse est correcte bien sûr). Attention cependant, dans ce cas, la régression nous donnera comme pente -b et comme ordonnée à l'origine ln a. On construit alors une ligne supplémentaire au tableau, en divisant la concentration de diiode par une concentration de référence C_0=1mol.L^{-1} afin d'assurer l'homogénéité

$\ln([I_2]/C_0)$	3,00	2,48	2,30	1,69	1,03	0,34	-0,29	-0,92	-2,30

On effectue alors la régression entre cette ligne et celle du temps. **N'hésitez pas à ajouter cette ligne supplémentaire lorsque vous effectuez une régression !** En cas d'erreur, vous trouverez plus facilement si vous avez gardé une trace de ce calcul intermédiaire.

On trouve effectivement une droite de pente

$$\ln [I_2] = 2{,}92 - 0{,}21.t$$

soit a=18,5 et b=0,21. La qualité est correcte car r=0,998 (on a deux 9 après la virgule).

Erreurs classiques

- Attention à l'erreur classique qui est l'échange de liste (la régression ne s'effectuant pas dans le bon ordre alors, voire entre deux listes qui ne correspondent pas à ce que vous voulez).
- Il faut TOUJOURS donner la valeur de r ou r^2, car sinon on ne justifie en rien le fait que la droite soit valable ou pas. Ne donnez pas trop de CS non plus, n'oubliez pas que l'on cherche approximativement une droite qui passe par des points expérimentaux ... ca fait beaucoup pour donner 6 CS tout ça !
- Si vous chercher à faire une régression sur une relation non linéaire, prenez le temps de faire apparaître les transformations qui font apparaître une loi linéaire, et de construire les lignes donnant la ou les deux nouvelles séries sur lesquelles vous allez faire votre régression. En voulant aller trop vite, vous pourriez rater la forme que vous cherchez.